

Математические методы в биологии

Блок 3. Математическая статистика

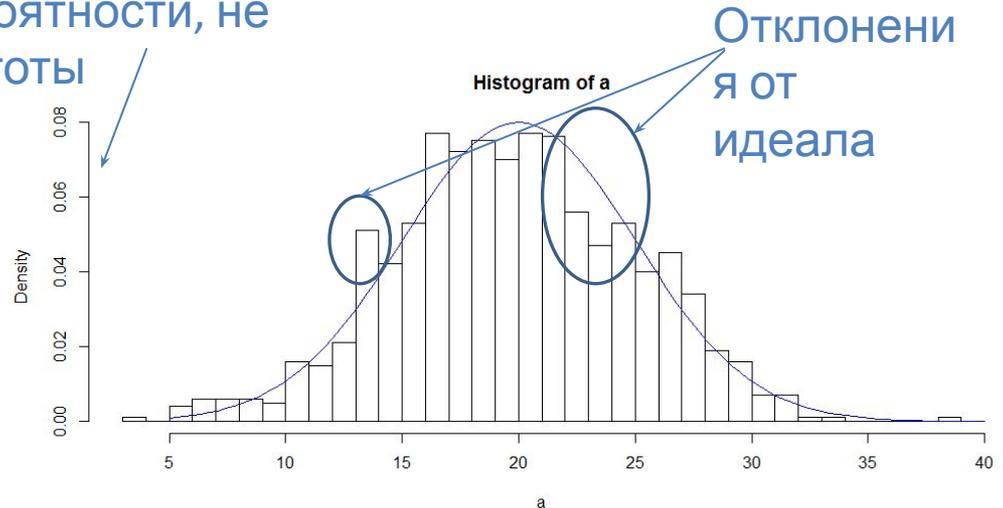
Лекция 6

Козлова Ольга Сергеевна
89276755130, olga-sphinx@yandex.ru

Проверка распределения на нормальность

1000 случайных значений,
распределённых по нормальному
закону с $\mu=20$ и $\sigma=5$
Синяя линия – кривая плотности
идеального нормального
распределения с $\mu=20$ и $\sigma=5$

Вероятности, не
частоты



Отклонени
я от
идеала

Любые экспериментальные данные всегда отклоняются от «сферического нормального распределения в вакууме»!

Выборочные
значения

Sample Quantiles

Normal Q-Q Plot

Значений здесь больше, чем должно быть для
н.р.

Значений здесь меньше, чем
должно быть для н.р.

Середина
распределения

Quantile-Quantile plot (Q-Q Plot)
Квантиль – значение,
которое делит
упорядоченную выборку на
несколько равных частей

Theoretical Quantiles

Предсказанные значения по норм.

Формальные тесты на нормальность

- Визуализация (гистограмма или Q-Q plot) позволяют определить, в каких конкретно точках выборочные значения отклоняются от нормального распределения. При этом Q-Q plot предпочтительней, когда наблюдений мало.
- Формальные тесты отвечают на вопрос, нормально ли распределение в принципе.

Тест Шапиро-Уилкса

H_0 : выборка распределена по нормальному закону (😊)

H_1 : выборка распределена по нормальному закону (😞)

⇒ Если $p\text{-value} > 0,05$ – распределение соответствует нормальному закону (😊)

Тест Колмогорова-Смирнова

H_0 : случайная величина X (значения признака в выборке) имеет распределение $F(X)$ (нормальное распределение – частный случай)

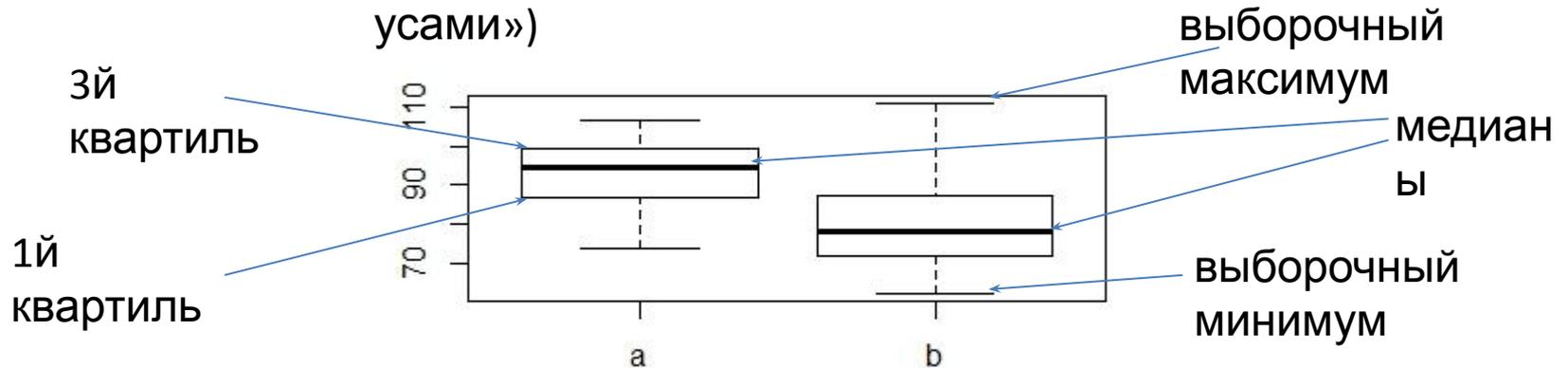
H_1 : её распределение отличается от $F(X)$

⇒ Если $p\text{-value} > 0,05$ – случайная величина имеет распределение $F(X)$

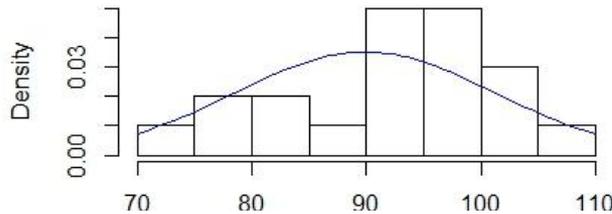
Почему это важно?

- Две нормальные выборки: $a(n=20, \mu=89.9, \sigma=11.3)$ и $b(n=20, \mu=80.7, \sigma=11.7)$

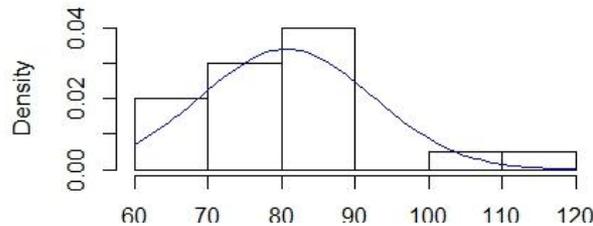
Диаграмма типа boxplot («ящик с усами»)



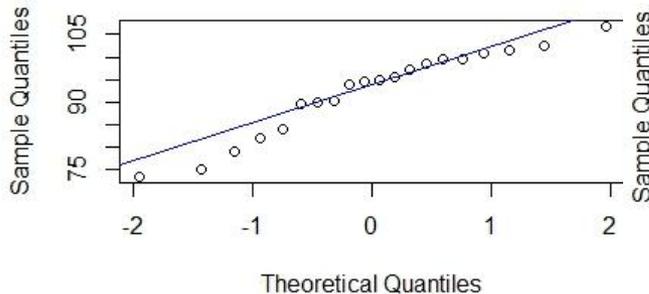
Histogram of a



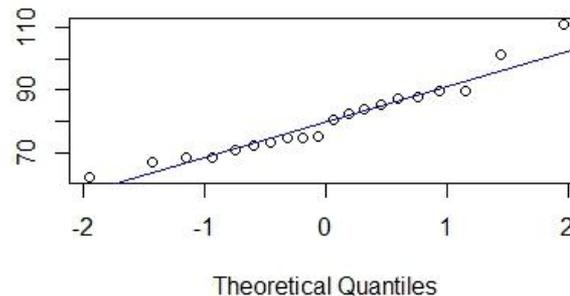
Histogram of b



Normal Q-Q Plot



Normal Q-Q Plot



Формальные тесты:

- Шапиро-Уилкс
 $p\text{-value}(a)=0,1722$
 $p\text{-value}(b)=0,2233$
- Колмогоров-Смирнов
 $p\text{-value}(a)=0,1626$

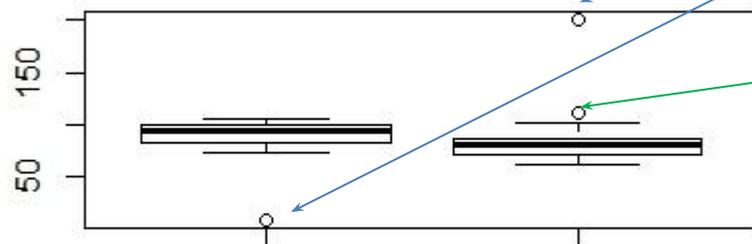
$p\text{-value}(b)=0,1595$

Тест Стьюдента:
 $p\text{-value} = 0,00112$
 $\Rightarrow H_0$ отвергаем
 средние не

равны!

Как испортить себе жизнь нормальность?

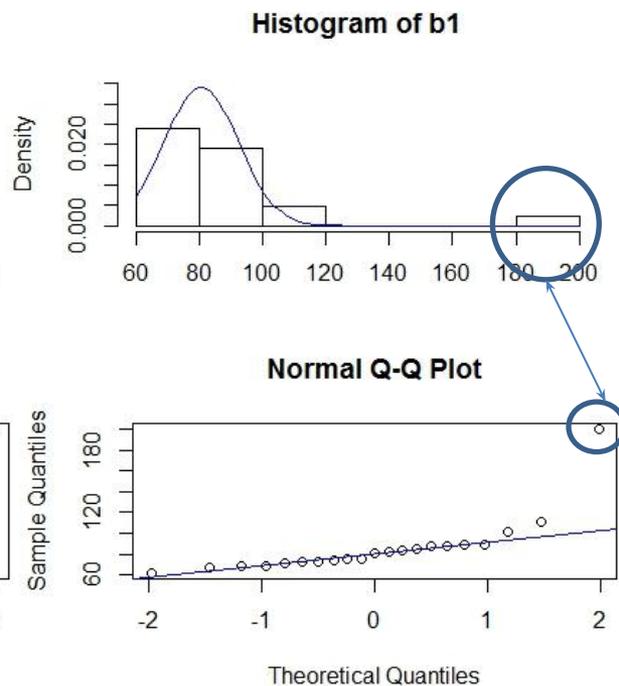
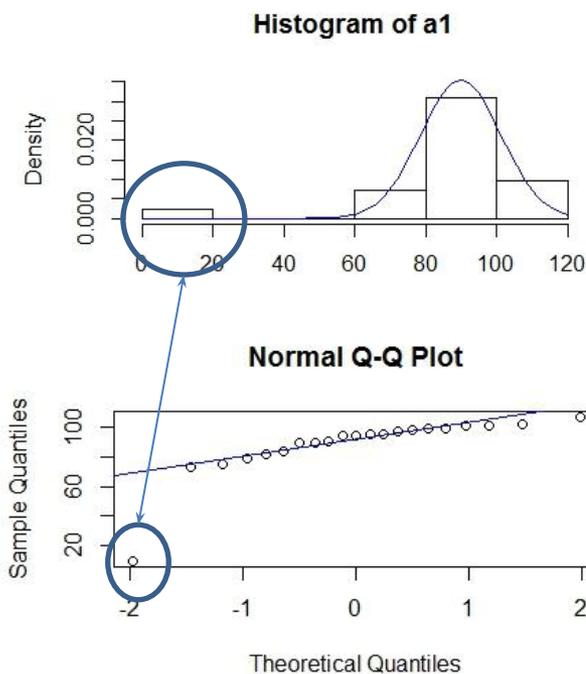
- Добавим экстремально отстоящие от выборки значения (выбросы)



Ещё один выброс, образованный вследствие сдвига квартилей

Формальные тесты:

- Шапиро-Уилкс
 - $p\text{-value}(a) = 6.725 \cdot 10^{-6}$
 - $p\text{-value}(b) = 2.202 \cdot 10^{-6}$
- Колмогоров-Смирнов
 - $p\text{-value}(a) = 0,003918$
 - $p\text{-value}(b) = 1.653 \cdot 10^{-5}$



Тест Стьюдента:
 $p\text{-value} = 0,7435$
 $\Rightarrow H_0$ не отвергаем

Непарам.аналог:
 $p\text{-value} = 0,01167$

Однофакторный дисперсионный анализ

- Сравниваем между собой не две, а несколько групп

Пример. Длина лепестка у ирисов трёх сортов

- Наблюдения делятся на группы по факторному (номинативному) признаку, выраженному независимой переменной

*Пример. Все собранные ирисы делятся на три группы – сорт *Versicolor*, сорт *Virginica* и сорт *Setosa*. Переменная «сорт ириса» независимая переменная.*

- Изучаем зависимую переменную – количественную переменную, выраженность которой зависит от независимой.

Пример. Зависимая переменная – длина лепестка ириса.



Virginica



Setosa



Versicolor

Задача: зависит ли длина лепестка ириса от того, к какому сорту он принадлежит?

Условный пример

Пусть собрано 9 цветков ириса – по 3 для каждого сорта.

Сорт ириса	Setosa	Virginica	Versicolor
Первый цветок (длина лепестка)	3	5	7
Второй цветок (длина лепестка)	1	3	6
Третий цветок (длина лепестка)	2	4	5

$H_0: \mu_{setosa} = \mu_{virginica} = \mu_{versicolor}$ (все выборки – из одной ГС)

H_1 : хотя бы одно истинное среднее отлично от остальных

Решение. Рассчитаем общее среднее для всех выборок:

$$\bar{x} = \frac{3 + 1 + 2 + 5 + 3 + 4 + 7 + 6 + 5}{9} = 4$$

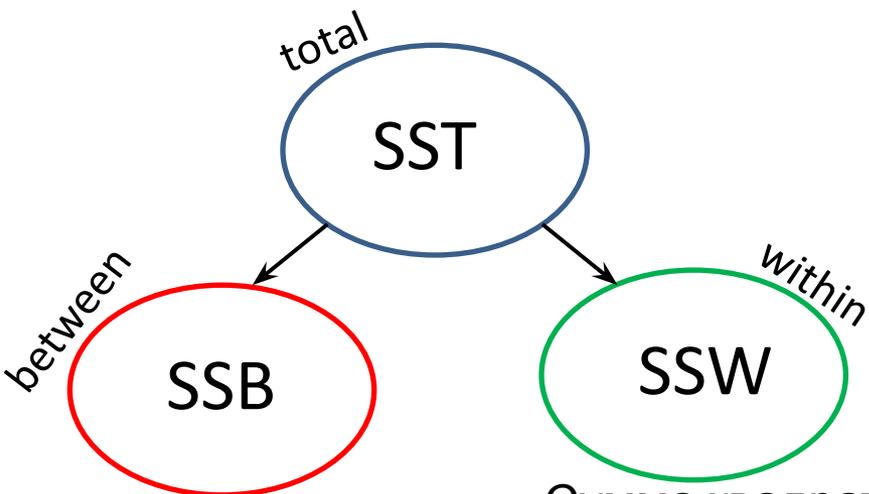
Введём понятие общей суммы квадратов отклонений (SST = sum of squares total). Это показатель, характеризующий изменчивость данных без учёта деления на группы.

$$\begin{aligned} SST &= \sum (x - \bar{x})^2 = (3 - 4)^2 + (1 - 4)^2 + (2 - 4)^2 + (5 - 4)^2 + (3 - 4)^2 + (4 - 4)^2 + (7 - 4)^2 + (6 - 4)^2 + (5 - 4)^2 \\ &= 1 + 9 + 4 + 1 + 1 + 0 + 9 + 4 + 1 = 30 \end{aligned}$$

Ещё об общей сумме квадратов

$$\bar{x}_1 = 2 \quad \bar{x}_2 = 4 \quad \bar{x}_3 = 6$$

Setosa	Virginica	Versicolor
3	5	7
1	3	6
2	4	5



Сумма квадратов отклонений между группами

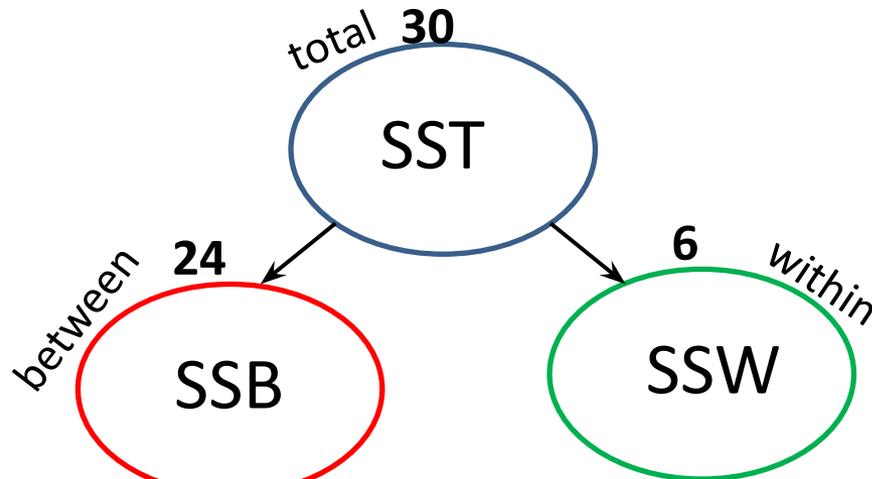
Сумма квадратов отклонений внутри групп

$$SSW = \sum_{\text{по 1й группе}} (x - \bar{x})^2 + \sum_{\text{по 2й группе}} (x - \bar{x})^2 + \sum_{\text{по 3й группе}} (x - \bar{x})^2 + \dots = (3 - 2)^2 + (1 - 2)^2 + (2 - 2)^2 + (5 - 4)^2 + (3 - 4)^2 + (4 - 4)^2 + (7 - 6)^2 + (6 - 6)^2 + (5 - 6)^2 = 6$$

$$SSB = \sum n(\bar{x} - \bar{\bar{x}})^2 = 3 * (2 - 4)^2 + 3 * (4 - 4)^2 + 3 * (6 - 4)^2 = 24$$

число элементов в группе

Итак,



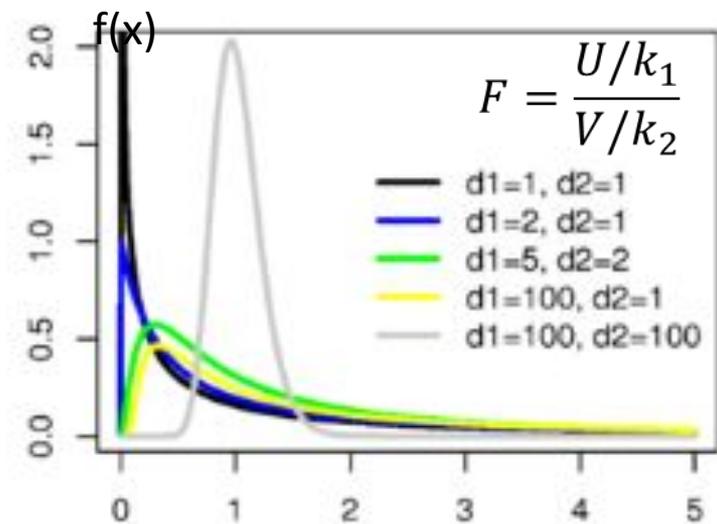
Сумма квадратов отклонений между группами

Сумма квадратов отклонений внутри групп

Большая часть общей изменчивости обеспечивается изменчивостью между группами, значит, группы-таки различаются между собой

Назад, к статистике: SSB и SSW – это случайные величины, имеющие распределение χ^2 (представляют собой суммы квадратов нормальных с.в.). Если скорректировать их на число степеней свободы и поделить SSB на SSW, получим с. в., распределённую по закону Фишера. Для SSB ч.с.св. = числу групп – 1, для SSW = числу наблюдений – число групп.

Плотность распределения



F-
значение

$$F = \frac{\frac{SSB}{m-1}}{\frac{SSW}{N-m}}$$

Смысл F-значения:
показывает, во сколько
раз межгрупповая
вариабельность
превышает
внутригрупповую

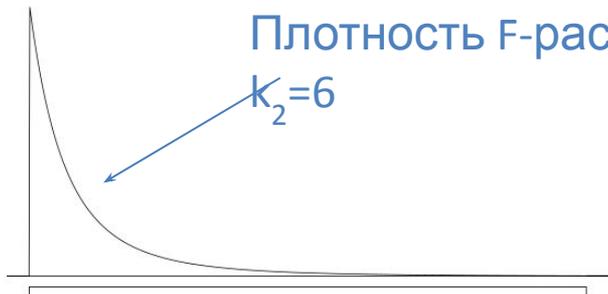
Число
групп

Число наблюдений (в
целом)

F-значение для нашего примера: $F = \frac{\frac{24}{2}}{\frac{6}{6}} = 12$

ВОПРОС: какова вероятность получить такое или ещё более крупное значение при условии, что верна H_0 ?

F Distribution



$P(X > 12) = 0.008$

Вероятность того, что межгрупповая вариабельность будет превышать внутригрупповую в 12 и более раз при условии равенства средних

Задача

- Будем изучать влияние генной терапии (независимая переменная) на уровень экспрессии гена (зависимая переменная).

Терапия	Объём выборки (число пациентов)	Выборочные средний уровень экспрессии	Выборочное стандартное отклонение
A	15	99,7	4,1
B	15	98,8	5,8
C	15	94,4	5,1
D	15	92,3	3,8

$$H_0: \mu_A = \mu_B = \mu_C = \mu_D$$

H_1 : хотя бы одна из генных терапий приводит к отличному от остальных уровню экспрессии

На компьютере (R):

Независ.
пер.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Therapy	3	560.7	186.91	8.037	0.000152
Residuals	56	1302.3	23.25		

$m-1$ (points to 3)
 SSB (points to 560.7)
 $SSB/(m-1)$ (points to 186.91)
 F -значение (points to 8.037)
 Ур-нь значимости (points to 0.000152)
 $N-m$ (points to 56)
 SSW (points to 1302.3)
 $SSW/(N-m)$ (points to 23.25)

Множественные сравнения

ВОПРОС: Можем ли мы теперь сказать, какая конкретно пара терапий статистически значимо различается между собой по уровню экспрессии гена?

ВСПОМНИМ КОМБИНАТОРИКУ И ТЕРВЕР. Сколько попарных сравнений надо выполнить, чтобы перебрать все возможные пары A,B,C,D?

$$C_4^2 = \frac{4!}{2!(4-2)!} = 6$$

Пусть пороговое значение отвержения $H_0 = 0,05$, т.е. вероятность совершить ошибку 1го рода для каждого из 6 тестов равна 0,05. Какова вероятность того, что хотя бы в одном из 6 тестов будет совершена ошибка первого рода (H_0 будет отклонена неправомерно, а различия между средними на самом деле случайны)?

Вероятность не-совершения ошибки = 0,95. $0,95^6 = 0,735$ (вероятность того, что не будет совершена ошибка в 6 тестах), значит, вероятность того, что хотя бы в одном из тестов она будет совершена, = $1-0,735=0,265$.

Вывод. Даже если различий между средними на самом деле нет, в 26,5% случаев при извлечении 4х выборок из одной ГС между какими-то из них мы будем получать статистически значимые различия!

Что же делать?

- Поправка на множественное сравнение Бонферрони.

Идея. Вероятность совершения ошибки первого рода растёт пропорционально увеличению числа попарных сравнений. Почему бы не уравновесить этот рост с помощью корректировки критического p-value в сторону убывания?

А именно, разделим критическое значение p-value на число попарных сравнений: $0,05/6=0,008333$. Тогда вероятность того, что в 6ти тестах будет совершена хотя бы одна ошибка 1го рода = $1-(1-0,008333)^6 = 0,049$.

НО! Сильное снижение критического уровня p-value ведёт к увеличению вероятности совершить ошибку 2го рода (H_0 не отвергается, хотя должна была бы).

- Альтернатива – использование критерия Тьюки (*критерий достоверно значимой разности Тьюки, Tukey's honestly significant difference test, Tukey's HSD test*)
 - похож на критерий Стьюдента, но стандартная ошибка среднего рассчитывается по-другому

Критерий Тьюки

- Пусть есть m групп: A, B, C, ...
- $H_0: \mu_A = \mu_B$, $H_1: \mu_A \neq \mu_B$

Для каждого из попарных сравнений рассчитывается величина:

$$q = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{SSW}{N - m} \cdot \frac{1}{n}}}$$

средняя внутригрупповая сумма квадратов

Число наблюдений в группе A (B)

Если число наблюдений в A и B разное, то

$$q = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{SSW}{N - m} * \frac{1}{2} * \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

В нашем примере

\$Therapy	разность средних diff	ДОВ.ИНТ-л	ур-нь значимости p adj	Стат.значимые рез-ты (дов.инт-л не включает 0)
		lwr	upr	
B-A	-0.9333333	-5.595897	3.7292308	0.9514195
C-A	-5.3333333	-9.995897	-0.6707692	0.0188860
D-A	-7.4000000	-12.062564	-2.7374358	0.0005424
C-B	-4.4000000	-9.062564	0.2625642	0.0710413
D-B	-6.4666667	-11.129231	-1.8041025	0.0029443
D-C	-2.0666667	-6.729231	2.5958975	0.6457935

Двухфакторный дисперсионный анализ

- Не одна независимая переменная, а две.

Пример. Уровень экспрессии гена в зависимости от дозировки лекарств (высокая/низкая) и возраста пациента (молодой/пожилой).

Возраст	Дозировка		sd	n
молодые	высокая	104,8	5,8	16
молодые	низкая	105,5	4,4	16
пожилые	высокая	101	5,1	16
пожилые	низкая	102,3	5,1	16

Результат дисперсионного анализа:

```
      Df Sum Sq Mean Sq F value Pr(>F)
age    1  197.5  197.45    7.570  0.0078 **
dose   1   16.9   16.91    0.648  0.4238
Residuals 61 1591.2  26.08
```

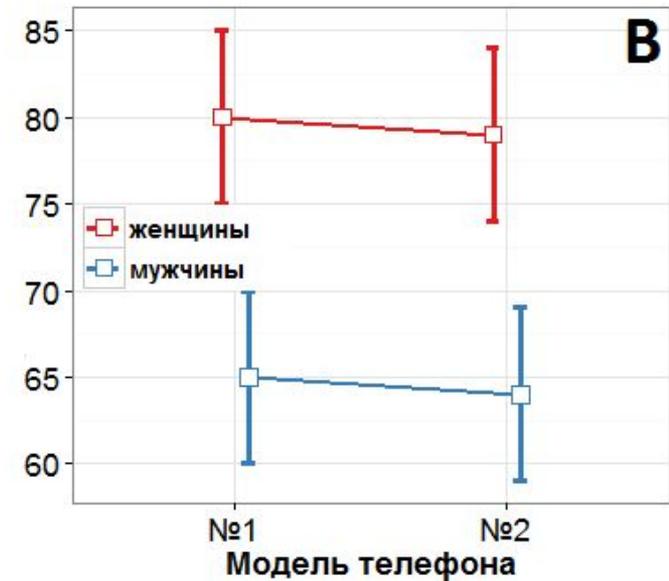
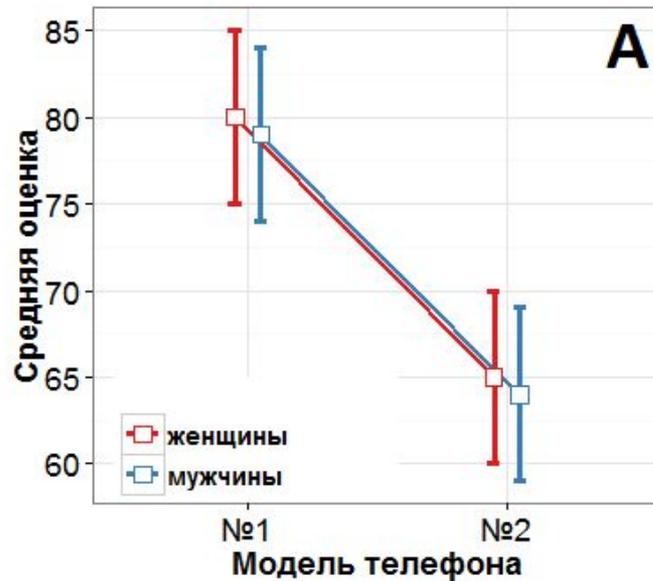
signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

В отличие от однофакторного, $SST = SSW + SSB_A + SSB_B + SSB_{A+B}$

Изменчивость,
обусловленная
взаимодействием
факторов

Как это выглядит?

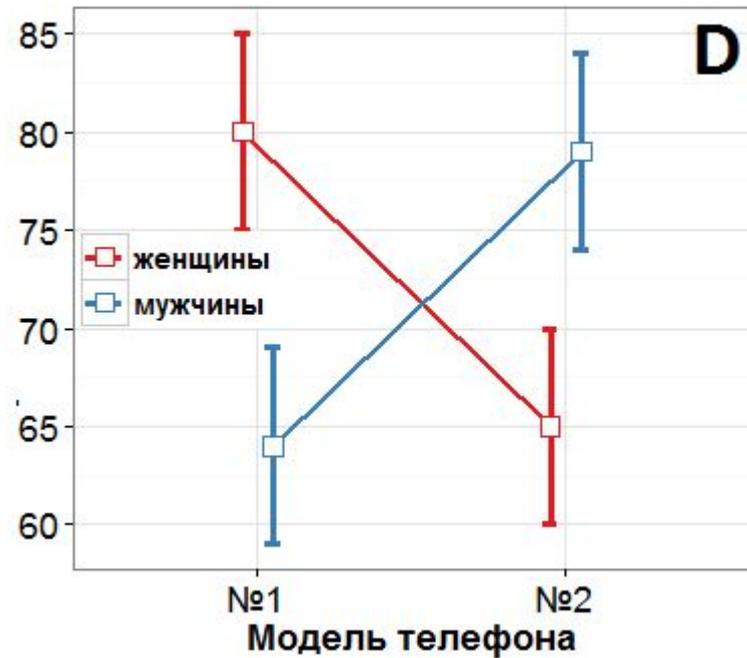
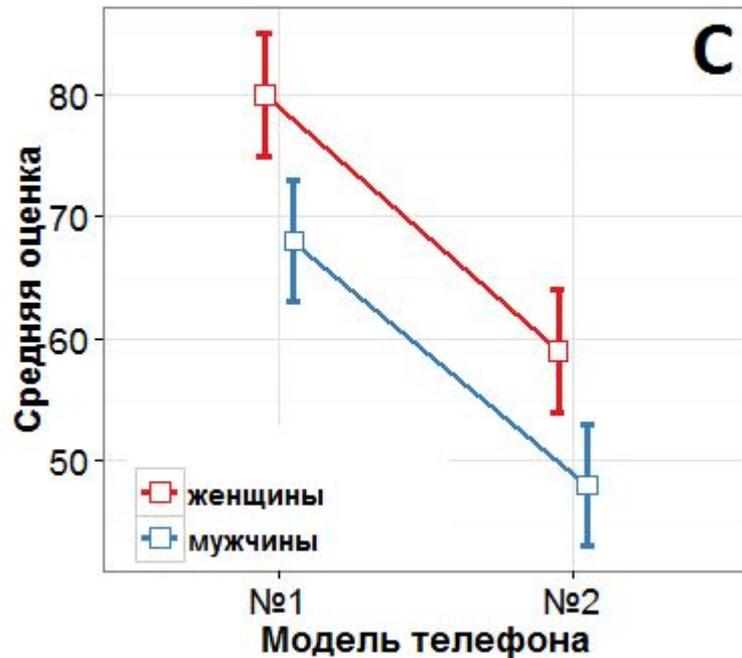
Фокус-группа из 100 мужчин и 100 женщин оценивает два телефона (модель №1 и модель №2) по 100-балльной шкале. Независимые переменные (факторы) – пол и модель телефона, зависимая переменная – оценка телефона по 100-балльной шкале.



А – значимый эффект только фактора «модель телефона» (и М, и Ж больше нравится 1я модель)

В – значимый эффект только фактора пола (женщинам в принципе больше нравятся телефоны)

Как это выглядит?



С – значимый эффект обоих факторов (женщинам в принципе больше, чем мужчинам, нравятся телефоны, но при этом и те, и другие более высоко оценили 1ю модель)

Д – значимое взаимодействие факторов (мужчины оценили вторую модель выше, чем первую, а женщины – наоборот). **Т.е. влияние одного фактора на зависимую переменную проявляется по-разному в зависимости от того, какое значение примет другая**

Требования к использованию дисперсионного анализа

- Нормальность распределения зависимой переменной в каждой из групп
- Гомогенность дисперсий (дисперсии признака внутри групп равны между собой)

Могут нарушаться при большом объёме выборок (>50).

Нормальность распределения проверяется:

- Графически (гистограмма плотности вероятностей, qq-plot)
- Формальными тестами (Шапиро-Уилкса, Колмогорова-Смирнова)

Гомогенность дисперсий проверяется:

- Графически (боксплот)
- Формальными тестами (тест Левена, при $p > 0,05$ дисперсии одинаковы)

Резюме по сравнению средних

- Для сравнения средних значений в двух группах – t-test
- Для сравнения средних в трёх и более группах – дисперсионный анализ
- Если результаты дисперсионного анализа говорят, что по крайней мере в двух группах средние различны, – использовать критерий Тьюки

Домашнее задание

- Посмотреть научно-популярный доклад «Статистика и плохая наука: как поправка на множественные сравнения объясняет парадоксальные результаты исследований»

Ссылка:

<https://www.youtube.com/watch?v=dcVG0NtZMwE>